Semi-Supervised Cause Identification from Aviation Safety Reports

Isaac Persing and Vincent Ng

The Aviation Safety Reporting System (ASRS)

- Established in 1967.
- ASRS collects aviation safety incident reports submitted by aviation personnel.
- As of July 2009, ASRS has over 150,000 reports available for download from http://asrs.arc.nasa.gov/ for use by researchers.
- Each report may contain up to 66 fields.

Cause Identification

- Each incident may be caused by one or many factors.
- For each test narrative, we want to find the shaping factors that influenced the incident it describes.

Shaping Factors

Id	Shaping Factor	Description
1	**Attitude**	Any indication of unprofessional or antagonistic attitude by a controller or flight crew member, e.g., complacency or get-homeitis (in a hurry to get home).
2	**Communication Environment**	Interferences with communications in the cockpit such as noise, auditory interference, radio frequency congestion, or language barrier.
3	**Duty Cycle**	A strong indication of an unusual working period, e.g., a long day, flying very late at night, exceeding duty time regulations, having short and inadequate rest periods.
4	**Familiarity**	A lack of factual knowledge, such as new to or unfamiliar with company, airport, or aircraft.
5	**Illusion**	Bright lights that cause something to blend in, black hole, white out, sloping terrain, etc.
6	**Other**	Anything else that could be a shaper, such as shift change, passenger discomfort, or disorientation.
7	**Physical Environment**	Unusual physical conditions that could impair flying or make things difficult.
8	**Physical Factors**	Pilot ailment that could impair flying or make things more difficult, such as being tired, drugged, incapacitated, suffering from vertigo, illness, dizziness, hypoxia, nausea, loss of sight or hearing.
9	**Preoccupation**	A preoccupation, distraction, or division of attention that creates a deficit in performance, such as being preoccupied, busy (doing something else), or distracted.
10	**Pressure**	Psychological pressure, such as feeling intimidated, pressured, or being low on fuel.
11	**Proficiency**	A general deficit in capabilities, such as inexperience, lack of training, not qualified, or not current.
12	**Resource Deficiency**	Absence, insufficient number, or poor quality of a resource, such as overworked or unavailable controller, insufficient or out-of-date chart, malfunctioning or inoperative or missing equipment.
13	**Taskload**	Indicators of a heavy workload or many tasks at once, such as short-handed crew.
14	**Unexpected**	Something sudden and surprising that is not expected.

Cause Identification Difficulties

- Narratives contain domain specific abbreviations and acronyms.

Decoding

- We expanded abbreviations using http://akama.arc.nasa.gov/ASRSDBOnline/pdf/ASRS_Decode.pdf .

- Original Sentence:

“HAD BEEN CLRED FOR APCH BY ZOA AND HAD BEEN HANDED OFF TO SANTA ROSA TWR.”

- After Decoding

“HAD BEEN CLEARED FOR APPROACH BY ZOA AND HAD BEEN HANDED OFF TO SANTA ROSA TOWER.”

Cause Identification Difficulties

- Narratives contain domain specific abbreviations and acronyms.
- Narratives have capitalization information removed.

Partial Case Restoration

- Heuristic case restoration using English lexicon.

- Original Sentence:

“HAD BEEN CLEARED FOR APPROACH BY ZOA AND HAD BEEN HANDED OFF TO SANTA ROSA TOWER.”

- Capitalization Restored Sentence

“had been cleared for approach by ZOA and had been handed off to santa rosa tower.”

Cause Identification Difficulties

- Narratives contain domain specific abbreviations and acronyms.
- Narratives have capitalization information removed.
- Cause Identification is a multi-class, multi-label problem.

Multi-class Multi-label Problem

- We treat cause identification as 14 independent binary classification tasks.

- Sample Narrative Extract:

"...I pressed on further and higher due to a belief the weather was not as bad as it was; and inexperience. …"

- Attitude
- Duty Cycle
- Illusion
+ Physical Environment
- Preoccupation
+ Proficiency
- Taskload
- Communication Environment
- Familiarity
- Other
- Physical Factors
- Pressure
- Resource Deficiency
- Unexpected

Cause Identification Difficulties

- Narratives contain domain specific abbreviations and acronyms.
- Narratives have capitalization information removed.
- Cause Identification is a multi-class, multi-label problem.
- Shaper-labeled narratives are scarce.

Data Scarcity

- We have only 1,333 shaper-labeled narratives.
- Some shapers are minority classes, each accounting for less than 10% of the labels applied to narratives.
- We can address these problems by automatically labeling some additional narratives from the 140,599 narrative corpus for use in training.

Bootstrapping Algorithm

- For each shaping factor S, begin with a labeled training set.
- If there are fewer positively labeled narratives in the training set than negativelly labeled ones, choose four words from the documents that are highly-positively correlated with S and add them to a set of positively correlated words.
- Add to the training set any narrative from the large unlabeled set containing at least three strongly correlated words.
- Iterate.

Bootstrapping Algorithm

$Train(P, N, U, k)$
Inputs:
P: positively labeled training examples of shaper x
N: negatively labeled training examples of shaper x
U: set of unlabeled narratives in corpus
k: number of bootstrapping iterations
$PW \leftarrow \emptyset$
$NW \leftarrow \emptyset$
for i = 0 to $k-1$ **do**
  if $|P| > |N|$ **then**
    $[P, PW] \leftarrow ExpandTrainingSet(P, N, U, PW)$
  else
    $[N, NW] \leftarrow ExpandTrainingSet(N, P, U, NW)$
  end if
end for

$ExpandTrainingSet(A, B, U, W)$
Inputs:
A, B, U: narrative sets
W: unigram feature set
for j = 1 to 4 **do**
  $t \leftarrow \arg\max_{t \notin W} \left(\log(\frac{C(t,A)}{C(t,B)+1})\right)$
  // $C(t, X)$: number of narratives in X containing t
  $W \leftarrow W \cup \{t\}$
end for
return $[A \cup S(W, U), W]$
// $S(W, U)$: narratives in U containing ≥ 3 words in W

Bootstrapping Algorithm Problems

- No way to recover from mislabeled narratives.

- What happens if we add to the positive training set all unlabeled narratives having only 1 of the positively correlated words?
  - The number of positively-correlated words a narrative from the unlabeled narrative set contains can be viewed as the amount of evidence we have that it should indeed be classified positively.
  - By adding narratives to the training set based on less evidence, we are likely introducing more noise into the training set.
  - We therefore require that any narrative we add to the training set contains at least 3 highly correlated (or highly negatively correlated) words. This still will not let us recover from the introduction of mislabeled narratives, but it makes recovery less necessary by slowing the introduction of noise into the training set.

Bootstrapping Algorithm Problems

- No way to recover from mislabeled narratives.
- Can get too specific to one subcategory of shaping factors.

Example from Physical Environment

- What happens if we select only one expansion word at a time?
- The word we select may deal specifically with only one subcategory of the shaping factor. For Example:
  - In one experiment, on the first iteration of the algorithm, we selected "snow" as an expansion word for Physical Environment.
  - On the next iteration, the word the algorithm selected was "plow".
  - Plow is highly positively correlated for Physical Environment, but its selection tells us snow-related narratives are now overrepresented in the training set.

Some Sample Expansion Words

Shaping Factor	Positive Expanders	Negative Expanders
Familiarity	unfamiliar, layout, unfamilarity, rely	
Physical Environment	cloud, snow, ice, wind	
Physical Factors	fatigue, tire, night, rest, hotel, awake, sleep, sick	declare, emergency, advisory, separation
Preoccupation	distract, preoccupied, awareness, situational, task, interrupt, focus, eye, configure, sleep	declare, ice, snow, crash, fire, rescue, anti, smoke
Pressure	bad, decision, extend, fuel, calculate, reserve, diversion, alternate	

- Most positive expanders make intuitive sense.
- Negative expanders make less intuitive sense.
- Some words appear as expanders for more than one set.

Baseline Classifiers

- We split the Cause Identification Task into 14 separate binary classification tasks, one for each shaping factor.
- We therefore construct 14 binary SVM classifiers using LIBSVM's probability option. A baseline classifier consists of a LIBSVM classifier for one shaping factor and a classification threshold between 0.0 and 1.0.
- For each narrative in a training set, we create 14 training examples, one for each classifier. Each example consists of a narrative's most relevant unigram features and the narrative's label with respect to this classifier's shaping factor.
- We say a baseline classifier labels a test narrative as positive if LIBSVM assigns it a probability above a given classification threshold.

Example Narrative Classification

Shaping Factor	LIBSVM Probability	Classification Threshold	Actual Label	Assigned Label	TP/TN/ FP/FN
1	0.11	0.25	-	-	TN
2	0.56	0.50	+	+	TP
3	0.42	0.35	-	+	FP
4	0.79	0.95	+	-	FN
…	…	…	…	…	…

Calculating Results

- Let c_i be the classifier dealing with shaping factor i; tp_i be the number of test reports correctly labeled as positive by c_i; p_i be the total number of test reports labeled as positive by c_i; and n_i be the total number of test reports that belong to shaping factor i according to the gold standard (Actual Label).
- We calculate Precision (P), Recall (R), and F-measure (F) as follows:

$$\mathrm{P} = \frac{\sum_i tp_i}{\sum_i p_i}, \mathrm{R} = \frac{\sum_i tp_i}{\sum_i n_i}, \text{and } \mathrm{F} = \frac{2PR}{P+R}.$$

Calculating Results

- We use these formulas to evaluate two different tasks:
  - For the 14 shaper classification task, i may take on any value from 1 to 14.
  - For 10 (minority) shaper classification task, i may only take on values corresponding to minority shaping factors (Attitude, Communication Environment, Duty Cycle, Familiarity, Illusion, Physical Factors, Preoccupation, Pressure, Taskload, and Unexpected).

Baseline Classifiers

- We define two types of baseline classifiers.
- $B_{0.5}$ has all classification thresholds set to 0.5.
- B_{ct} has classification threshold parameters that are tunable. The classification threshold for each binary classifier may be set to a different value.

Bootstrapping Classifiers

- We define two types of classifiers using the training sets obtained with our bootstrapping algorithm.
- $E_{0.5}$ is based on $B_{0.5}$, but it has one tunable parameter, the number of iterations we apply our bootstrapping algorithm before using the resulting training set.
- E_{ct} is based on B_{ct}, but its iteration parameter is tunable alongside its classification threshold parameter.

Results

- 5-fold cross validation results:

System	All 14 Classes			10 Minority Classes		
	P	R	F	P	R	F
$B_{0.5}$	67.0	34.4	45.4	68.3	23.9	35.4
B_{ct}	47.4	59.2	52.7	47.8	34.3	39.9
$E_{0.5}$	60.9	40.4	48.6	53.2	35.3	42.4
E_{ct}	50.5	54.9	52.6	49.1	39.4	43.7

Conclusions

- For all types of classifiers, F-measure tends to be about 10% greater on the 14 shaper classification task. We believe the difference can be attributed to two factors:
  - There are more positive training examples of the majority classes, making those classification subtasks easier.
  - It is simply easier to get high F-scores on tasks where the label we're trying to predict has a high frequency.

Conclusions

- The F-scores of the ct classifiers are always better than the F-scores of the corresponding 0.5 classifiers. This illustrates the importance of the classification threshold, but we believe the main reason it helps might be trivial.
  - The ct classifiers allow us to choose how many narratives should be labeled positive for each shaping factor by tuning the classification threshold parameter. The simulated annealing algorithm for parameter tuning tries to maximize F-measure by finding the best balance between precision and recall. The 0.5 classifiers do not have much control over what balance between precision and recall they obtain.

Conclusions

- Our best classifier E_{ct}'s performance on the 14 shaper task is slightly worse than that of the best baseline B_{ct}. On the 10 minority shaper task, however, using E_{ct} yields a 6.3% relative error reduction in f-measure over B_{ct}. We believe this is attributable to two factors:
  - Some minority classes are so scarce that the addition of even relatively noisy training data can improve classifier performance.
  - Majority classes may have enough positive training examples in the initial training set to train a decent classifier. The addition of slightly noisy training data may just make correct classification harder.

Some Obvious Further Improvements

- We notice that some pairs of shaping factors are highly positively correlated, and some are highly negatively correlated.
  - In our paper, we split Cause Identification into 14 independent binary classification tasks. Future approaches might make use of the connections between shaping factors.
- The reports as we found them on the ASRS website contain many fields besides the narrative field.
  - In our paper, we looked only at the narrative field. Future approaches could make use of the information contained in the other fields.

Acknowledgements

We thank the three anonymous reviewers for their invaluable comments on an earlier draft of the paper. We are indebted to Muhammad Arshad Ul Abedin, who provided us with a preprocessed version of the ASRS corpus and, together with Marzia Murshed, annotated the 1,333 documents. This work was supported in part by NASA Grant NNX08AC35A and NSF Grant IIS-0812261.

Works Cited

- Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: A library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm
- Christian Posse, Brett Matzke, Catherine Anderson, Alan Brothers, Melissa Matzke, and Thomas Ferryman. 2005. Extracting information from narratives: An application to aviation safety reports. In *Proceedings of the Aerospace Conference 2005*, pages 3678–3690.